A Flexible Volumetric Comparison of Protein Cavities can Reveal Patterns in Ligand Binding Specificity

Ziyi Guo[*], Trevor Kuhlengel[*], Steven Stinson, Seth Blumenthal, Brian Y.Chen[†]
Department of Computer Science and Engineering,Lehigh University
Bethlehem,PA,18015
chen@cse.lehigh.com

Soutir Bandyopadhyay
Department of Mathematics,Lehigh University
Bethlehem,PA,18015
sob210@lehigh.com

ABSTRACT

Conformational flexibility is an underlying cause of error in all comparisons of protein structure. Using flexible representations, some comparison algorithms can identify subtle functional similarities among distantly related proteins even when they exhibit different backbone conformations. The same techniques are not designed to identify subtle variations among closely related proteins that might cause differences in specificity. In such cases, molecular flexibility obscures structural details that influence the specific recognition of similar but non-identical ligands.

To enhance the analysis of ligand binding specificity, this paper presents FAVA (Flexible Aggregate Volumetric Analysis), a conformationally robust tool for comparing similar binding cavities with different binding preferences. FAVA examines a large number of conformational samples to characterize local flexibility using Constructive Solid Geometry. Using molecular dynamics simulations as a source for conformational samples, we used FAVA to analyze a nonredundant sample of serine protease and enolase structures. Different snapshots from the same proteins exhibited significant variations in binding cavity shape. Nonetheless, analysis with FAVA revealed subfamilies with different binding preferences. FAVA also identified amino acids associated with differences in binding preferences, predicting established experimental results. These results illustrate a new approach to flexible comparison that uses sampled conformational data. It reveals that detailed comparisons of very similar proteins, such as those within small ligand binding cavities, are possible even in the presence of conformational flexibility. Identifying influences on specificity in this manner points to new applications of protein engineering and drug design.

[*]equal contribution.
[†]Corresponding author.

BCB '13, September 20 - 23, 2014, Newport Beach, CA
ACM 978-1-4503-2434-2/13/09.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms

Algorithms,Design

1. INTRODUCTION

Algorithms for comparing protein structures widely employ the simplification that proteins are rigid. Many comparisons achieve considerable efficiency because rigid transformations move atoms from different structures into superposition without a lengthy analysis of alternative conformations. This simplicity permits many techniques to rapidly align backbone carbons [30, 31, 36, 42, 33, 9, 7], distance matrices [21], or geometric graphs [34, 15, 47] to discover remote homologs and proteins with similar functional sites. The rigidity simplification is also essential for a different class of comparison methods that seek to distinguish between closely related proteins with different ligand binding preferences [13, 10, 7]. Without rigidity as a starting point, the comparative structure-based analysis of function and specificity would be considerably more difficult.

Comparisons that tolerate greater conformational generality apply specialized flexible representations. A recent class of algorithms employ hinges [40, 17], graph structures [49, 23], fragments [28], and dynamic programming [46, 6, 26] to represent proteins as collections of rigid components with flexible linkers. Most approaches reported to date focus on discovering remote homologs that might be overlooked because of different conformations. But conformational flexibility also hinders the comparative analysis of closely related proteins with different ligand binding preferences. In such cases, the proteins considered may be very closely related, but side chain or backbone flexibility may obscure similarities or variations at ligand binding cavities that affect specificity. To address these issues, this paper focuses on the flexible comparison of ligand binding cavities, based on a survey of sampled conformations, to predict differences in specificity in spite of flexible variation.

The problem we are specifically addressing is the case where conformational samples of two or more proteins are available, and it is of interest to identify conserved or varying regions in their binding cavities that are preserved over many, but not necessarily all samples. Regions inside cavities that are solvent accessible in many samples of all proteins (*conserved frequent* regions) might accommodate a molecular fragment that is common to substrates acted on by all proteins (Fig. 1j). Alternatively, cavity regions that are solvent accessible in many samples of some proteins but

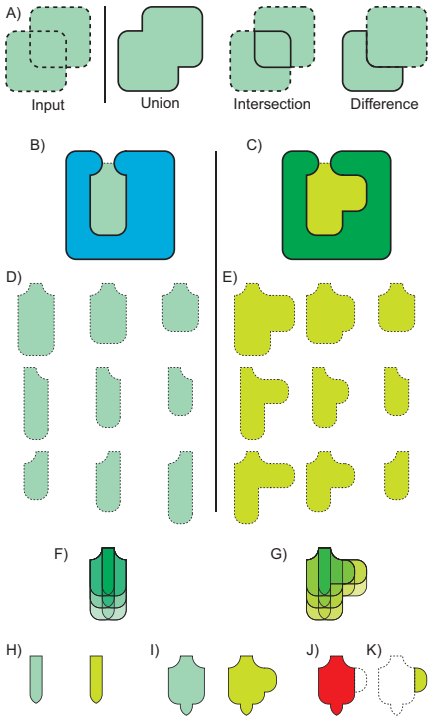

Figure 1: An overview of the FAVA method. A) CSG operations used by FAVA, with input regions (green, dotted outline) and output regions (solid outline). B,C) Input proteins X (blue) and Y (green) with ligand binding sites x (light green) and y (yellow). D,E) x and y in different conformational samples. F,G) All conformational samples of x (transparent green) and y (transparent yellow), superposed, with black outlines. Considerable variations in cavity shape are apparent. H) Using CSG, the intersection of all cavity regions in both proteins is too small to accommodate ligands. They are also identical, revealing little about different binding preferences. I) Using CSG, FAVA approximates frequent regions, where every point is inside at least two thirds of all samples of x (green) and y (yellow). J) The intersection of frequent regions indicates regions that might accommodate similar molecular fragments (red). K) The difference between frequent regions (yellow) indicates a region where Y might often accommodate a ligand that X cannot, causing a difference in specificity.

not often accessible in others (*unconserved frequent* regions), might cause those proteins to prefer different substrates (Fig. 1k). We will identify regions like these using the new method FAVA (Flexible Aggregate Volumetric Analysis).

Our approach with FAVA is to represent the three dimensional region that is frequently, but not universally, within the ligand binding cavity of a protein. We call this region the *frequent region* (Fig. 1i) because it ignores the geometry of unusual conformations that can obfuscate the region that is typically solvent accessible in the cavity. FAVA computes frequent regions using *operations* from Constructive Solid Geometry (CSG) [10], such as union, intersection and difference (Fig. 1a). The same operations enable comparisons of frequent regions: intersections produce conserved frequent regions, and differences produce unconserved frequent regions. Finally, CSG operations permit the identification of amino acids that frequently alter cavity shape, and thus have a steric influence on specificity. Together, these techniques create a conformationally general approach for examining closely related proteins in search of influences on binding specificity.

Sampled conformations, as a flexible representation of protein structures explored first in this paper, exhibit novel advantages. First, all-atom samples provide a more general representation of molecular movement than rigid components. In every sample, the positions of every atom can be used to produce a detailed representation of ligand binding sites for aggregate comparisons, whereas rigid components and flexible linkers can create uncertainties in the positions of some atoms. Second, samples generated by molecular dynamics simulations, which we use in this work, are subject to the constraints imposed by biophysical energy functions. As a result, every sample represents a semi-realistic conformation, whereas existing flexible methods cannot make such guarantees. These advantages are acquired in exchange for the computational cost of simulation and of processing far more structural data for a comparison between a few proteins. For an analysis of ligand binding preferences where long timescale movements do not play a major role in binding, conformational samples may be an effective flexible representation for structure comparison.

We demonstrate the capabilities of FAVA on subfamilies of two protein superfamilies: the serine proteases and the enolases. Conformational samples were selected from 100 nanosecond simulations of 19 sequentially nonredundant representative structures, and cavities from each sample were analyzed with FAVA. In many cases, FAVA sensitively classified cavities into subfamilies with different binding preferences, despite the fact that many conformational samples exhibited degenerate or misleadingly shaped cavities. We also observed that FAVA was able to correctly identify amino acids that create steric influences on ligand binding in spite of sidechain flexibility. These capabilities point to novel applications in protein engineering and the characterization of ligand binding specificity in ways that mitigate inaccuracies from conformational noise.

One such application, for example, is in support of structural studies that seek to discover the elements of protein structures that cause related proteins to exhibit different binding preferences. In such cases, the flexibility of protein backbones and sidechains can obscure the underlying mechanisms in one binding site that stabilize ligands that are not accommodated in others. FAVA could be applied to identify the cavity regions or amino acids that influence specificity, potentially reducing the time consuming and expensive experimentation necessary to isolate them otherwise. FAVA can thus assist in the discovery of mutations that cause drug resistance, alter signal transduction, and otherwise reorganize molecular interactions in biological systems.

2. METHODS

Formally, we define a frequent region as the region in space that is solvent accessible in more than k/N samples, where k, the *overlap threshold*, is provided as input, and N is the number of samples. When $0 < k/N < 1$, frequent regions represent cavity regions that are solvent accessible in several conformational samples without being restricted to unusual conformations that occur less frequently than k/N. FAVA is

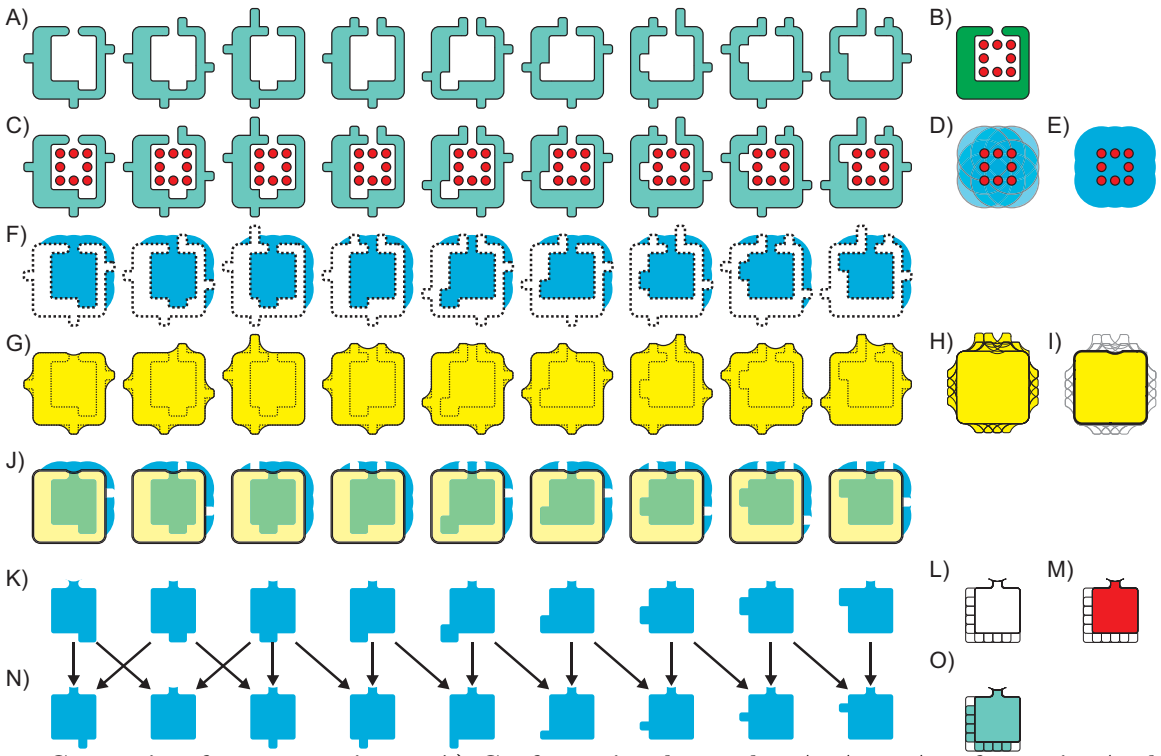

Figure 2: Generating frequent regions. A) Conformational samples $A_0, A_1, \ldots A_N$ of protein A, shown as molecular surfaces (teal, black outlines). B) Ligand l (red dots) bound to protein A (green). C) Aligning each A_i to A permits l to mark the ligand binding site in A_i. D) Spheres that define the neighborhood around the atoms of l (transparent blue, black outlines). E) CSG union of the spheres, S_l (blue). F) CSG difference A_i' that removes the molecular surface of each A_i (dotted outline) from copies of S_l. G) Envelope surfaces (yellow, black lines) and molecular surfaces (dotted lines) of all A_i. H) Envelope surfaces aligned. Outlines of all envelopes are shown in black. I) The global envelope region, $E(A)$, generated with CSG intersections (yellow, heavy black outline). J) CSG intersection between each A_i', shown in blue, and $E(A)$ (transparent yellow, black outline). K) Cavities a_i defined on each conformational sample. L) Cavity borders superposed (black outlines). M) frequent region that overlaps at least 3 cavities (red). N) CSG intersections between several pairs of a_i. O) The CSG union of intersections in N: $\alpha_k^\star$, the approximated frequent region (teal) overlapping at least 2 cavities.

completely agnostic as to the source of the conformational samples, which can be drawn from experimental or computational data, such as structural restraints from NMR, molecular dynamics trajectories, and others. Below, we describe how we compute frequent regions using a series of CSG operations, but we leave the description of individual CSG operations (union, $\cup$, intersection, $\cap$, and difference, $-$) to earlier publications [10]. We then describe how we compare frequent regions from multiple proteins to identify conserved frequent and unconserved frequent regions. Finally, we explain how we use solid representations to characterize the flexible geometry of individual amino acids and their steric impingement on nearby binding cavities.

2.1 Generating frequent regions

As input, we require the overlap threshold k, N conformational samples of a protein structure A, and a ligand l bound to A (Fig. 2b). We refer to the conformational samples as $A_0, A_1, \ldots A_N$. Samples can be provided from any source that specifies the position of every atom, with the expectation that sufficient samples are provided to describe short timescale motion near the binding cavity. From this data, generating a frequent region occurs abstractly in two steps, where we first define the shape of the ligand binding cavity in each conformational sample, and then use those cavities to determine the shape of the frequent region.

First, every sample A_i is superposed onto A by minimizing the root mean squared distance between identical amino acids [45]. Next, in every A_i, we use GRASP2 [33] to generate the molecular surface $m(A_i)$ (Fig. 2a). This surface is defined by the classical rolling probe algorithm [11] with the standard probe size of 1.4Å. Since every conformational sample is superposed onto A, we use l to locate the ligand binding site in every superposed $m(A_i)$ (Fig. 2c).

At every atom in l, we center a sphere with radius 5 Å (Fig. 2d). The CSG union of the spheres defines a neighborhood, S_l (Fig. 2e), that defines the vicinity of the ligand binding cavity in every sample. Making a copy of S_l for every A_i, we compute a CSG difference A_i' that removes the molecular surface of each A_i from the copy of S_l, revealing part of the cavity (Fig. 2f).

Next, we generate an *envelope surface*, $e(A_i)$, for every sample. The envelope surface is also generated with GRASP2, except that the probe radius is changed to 5.0Å (Fig. 2g). Because the larger probe is 10 Å in diameter, it does not roll into smaller clefts and cavities, making $e(A_i)$ a logical exterior boundary between the cavity and the solvent. Since $e(A_i)$ can vary significantly between different conformational samples, especially because of solvent-facing side

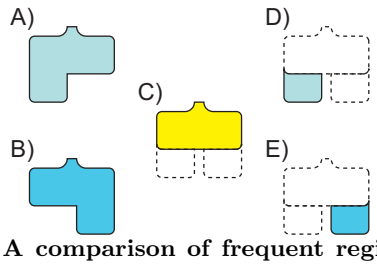

Figure 3: A comparison of frequent regions. A,B) Frequent regions $\alpha_k^\star$ (teal) and $\beta_k^\star$ (light blue). C) Conserved frequent region, $FC(A,B)$ (yellow). D,E) unconserved frequent regions (teal, light blue).

chains (Fig. 2g) unrelated to cavity shape, we mitigate these differences by computing their intersection (Fig. 2h,i):

$$E(A) = \bigcap_{\forall i} e(A_i) \tag{1}$$

We refer to $E(A)$ as the global envelope region. Because our samples are generated at a medium timescale, larger backbone motions are not a major factor in the shape of $E(A)$. Next, we compute the CSG intersection of A_i' and $E(A)$ (Fig. 2j). The result is the binding cavity in every conformational sample, $a_i = (S_l - m(A_i)) \cap E(A)$ (Fig. 2k).

We use the sampled cavities a_i together to approximate the frequent region α_k. Before we approximate this region, it is critical to recognize first that computing α_k explicitly, on a protein with many sampled conformations, is computationally impractical for many k. Consider, for example, the simple case of $k = 30$. The region α_{30} includes the CSG intersection of $a_0, a_1, a_2, \ldots, a_{30}$, because any point inside all of these regions is inside at least 30 a_i, and thus inside α_{30}. The same is true for any thirty member subset of $\{a_0, a_1, \ldots, a_N\}$, so α_{30} is the union of all intersections of thirty distinct sample cavities: $\binom{N}{30}$ intersections. Where N is several hundred samples and k is nontrivial, the exponential size of the calculation is clearly impractical, given the number of combinations.

FAVA approximates α_k by randomly selecting subsets of size k. We call the approximated result $\alpha_k^\star$, and compute it in the following manner: Given any k, we randomly select 500 distinct subsets of $\{a_0, a_1, \ldots, a_N\}$ of size k, and compute their CSG intersection (a smaller, diagrammatic random selection is shown in Fig. 2n). Finally, we compute the CSG union of the resulting intersections, $\alpha_k^\star$ (Fig. 2o). While random selections of different sizes were tested (see Supplementary materials, Section 1), frequent regions based on different random subsets of 500 had consistent volumes. We deemed 500 samples to be sufficient for accurate representations.

2.2 Comparing frequent regions

Given two proteins A and B, we use their frequent regions $\alpha_k^\star$ and $\beta_k^\star$, to evaluate the similarities and differences of their ligand binding sites over time (Fig. 3). These calculations are only performed once both structures and all conformational samples are structurally aligned, to avoid errors from poor superposition. Specifically, B is structurally aligned onto A using ska [48]. Next, every snapshot A_i is superposed onto A and every B_i is superposed onto B respectively, by minimizing the root mean squared distance between identical amino acids [45]. If more than two proteins were being considered, they would also be aligned to A first. Ultimately, every snapshot is aligned to A and B, which were first aligned onto each other.

After all superpositions are performed and frequent regions $\alpha_k^\star$ and $\beta_k^\star$ are calculated, we use the frequent regions to compute conserved frequent regions. The conserved frequent region between the samples of A and B is $FC(A,B) = \alpha_k^\star \cap \beta_k^\star$ (Fig. 3e). Because $FC(A,B)$ is the region conserved between two frequent regions, it approximates a binding cavity region that is solvent accessible in both proteins in more than k conformational samples. We measure the *volumetric distance*, $D(A,B)$, between the frequent regions of two proteins using the following expression,

$$D(A,B) = 1 - \frac{|FC(A,B)|}{|\alpha_k^\star \cup \beta_k^\star|}, \tag{2}$$

where the expression $|x|$ denotes the volume within a solid region x. We measure volumes using the Surveyor's formula [37], which we described earlier [10].

A comparison of cavities a_i and b_j from individual conformational samples of two different proteins is also possible. We evaluate their volumetric distance as:

$$d(a_i, b_j) = 1 - \frac{|a_i \cap b_j|}{|a_i \cup b_j|} \tag{3}$$

2.3 Frequently influential amino acids

Given two proteins A and B, if the cavity of A is frequently different from B, then some set of amino acids is responsible for making these cavities different on a frequent basis. We identify such amino acids with FAVA.

At the level of individual samples, consider two samples of A and B, called A_i and B_j, and an amino acid r in A. We say that r makes the cavity a_i different from the cavity b_j if the intersection of the molecular surface of r in A_i, called $m(r_i)$, has a nonempty intersection with b_j. If so, then $m(r_i)$ occupies a region that is not solvent accessible in a_i but solvent accessible in b_j. Between these two samples r_i is thus one cause for the difference between a_i and b_j.

To evaluate how frequently r, an amino acid of A, creates differences between the cavities of A and B, we compute $INT_r(A,B)$, the median volume of intersection $|m(r_i) \cap b_j|$, for all pairs of samples A_i and B_j. When $INT_r(A,B)$ is large, then r frequently makes the cavity of A different from B; small values indicate that it rarely does.

2.4 Data set construction

To demonstrate that FAVA can separate proteins with different binding preferences, we selected two superfamilies based on established results documenting the existence of distinct families in each superfamily with different binding preferences (Figure 2.4). Within the serine proteases, we selected the trypsin, chymotrypsin, and elastase subfamilies. In the enolase superfamily, we selected the enolase, mandelate racemase, and muconate lactonizing enzyme families.

Serine proteases selectively cleave peptide bonds using a nucleophilic serine residue. Preferences for hydrolyzing a specific scissile bond are achieved by recognizing amino acids on both sides of the bond, most notably the $P1$ residue immediately before the bond. The $S1$ specificity pocket, which recognizes $P1$, is large and hydrophobic in chymotrypsins and prefers to bind large hydrophobic residues [27]. In trypsins, $S1$ stabilizes positively charged amino acids, complementing its notable negative charge [16]. Enolases exhibit a

Serine Protease Superfamily:
Chymotrypsins: 1ex3
Elastases: 1b0e, 1elt
Trypsins: 1a0j, 1ane, 1aq7, 1bzx, 1fn8, 1h4w, 1trn, 2eek, 2f91
Enolase Superfamily:
Enolases: 1iyx, 1ebh, 1te6, 3otr
Mandelate Racemase: 1mdr, 2ox4
Muconate Lactonizing Enzyme: 2pgw

Figure 4: PDB codes of structures used.

small hydrophobic $S1$ cavity that binds small hydrophobic amino acids [4].

Members of the enolase superfamily exhibit a TIM-barrel fold and an N-terminal "capping domain" [35]. Using amino acids at the C-terminal ends of beta sheets in the TIM-barrel, superfamily members achieve a range of different functions that generally abstract a proton from a carbon adjacent to a carboxylic acid [1]. The enolase family catalyzes the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate [24], mandelate racemases convert (R)-mandelate to and from (S)-mandelate [38], and muconate lactonizing enzyme catalyze the reciprocal cycloisomerization of cis,cis-muconate and muconolactone.

Selection. Serine protease and enolase structures were selected from the protein data bank (PDB) [5] on 6.21.2011. Based on enzyme classifications (EC), the PDB contained 676 serine proteases and 66 enolases in the families selected for our data set. From these structures, mutants, structures with disordered regions, and enolases with closed or partially closed capping domains were removed. Next, one structure from any pair of structures with greater than 90% sequence identity was removed, with a preference for keeping structures associated with publications. Technical problems with simulation prevented proteins 8gch, 1aks, and 2zad from being added. From the 12 serine protease and 7 enolase structures that remained, ions, waters, and other non-protein atoms were removed. Hydrogens, unavailable in some structures, were also removed, but non-canonical amino acids (e.g. selenomethionines) were not removed.

Alignment. For the all-pairs comparison of frequent regions, we superposed all structures and conformational samples in each superfamily. All serine proteases and their samples were superposed onto 8gch, a chymotrypsin, and all enolases and enolase samples were superposed onto 1mdr. These structures were selected because of the presence of a bound ligand, which we used to define the binding cavity.

2.5 Protein structure simulation

For each structure in the data set, conformational samples were computed using GROMACS 4.5.4 [19]. To prepare for the simulation, a cubic waterbox is created, and the protein molecule is centered in the box. The box was populated using SPC/E, an equilibrated 3-point solvent model [3]. Fully periodic boundary conditions were used throughout the equilibration and simulation steps. The waterbox size was set to contain the solute protein structure with a 1.0 nanometer space between the protein and the nearest point on the boundary plane. Charge balanced sodium and potassium ions were then added to the solvent at a low concentration ($< 0.1\%$ salinity).

Energy minimization using a steepest descent algorithm is then performed for the entire system. Neutral Pressure Temperature (NPT) equilibration is performed in four 250 picosecond steps to allow the solvent to equilibrate temperature and pressure prior to the primary simulation. Starting at 1000 $kJ/(mol * nm)$, each step reduced the position restraint force by 250 $kJ/(mol * nm)$ over the 1 nanosecond minimization period. Backbone position restraints were released for the primary NPT simulation.

System energies were generated at the start of the equilibration phase. Initial temperature was 300 Kelvin and initial pressure was 1 bar. The Nosé-Hoover thermostat [3] was used for temperature coupling. The P-LINCS [18] bond constraint algorithm was used to update bonds. Electrostatic interaction energies were calculated by particle mesh Ewald summation (PME) [20]. The Parrinello-Rahman algorithm was used for pressure coupling [32, 29]. All temperature and pressure scaling was performed isotropically.

Full MD simulation is started using the atomic positions and velocities of the final equilibration state. The total simulated duration of the molecular dynamics simulation was 100 nanoseconds, with 1 femtosecond steps. P-LINCS and PME were chosen for their parallel efficiency. OpenMPI was used for node and inter-process communication. Simulations were run on multiple nodes with 16 cores each, with PME distribution automatically selected by GROMACS.

After simulations were completed, the trajectory file was converted to a simple Protein Data Bank format with atom positions only. The waterbox was removed and at each timestep, the protein was rigidly superposed to the original orientation. From these timesteps, we selected 600 conformational samples at uniform intervals for our data set, and then computed frequent regions.

2.6 Clustering frequent regions by volumetric distance

We hypothesize that similarities and differences between frequent regions can be used to classify samples of ligand binding cavities based on their binding preferences. To evaluate this hypothesis, we generated frequent regions with an overlap threshold of 50, and measured volumetric distance between all pairs of frequent regions in the same superfamily. We then used the neighbor tool from Phylip [14] to perform UPGMA clustering (Unweighted Pair Group Method with Arithmetic mean) [44] based on volumetric distance.

To demonstrate the one benefit of a flexible representation of binding cavities, we compared the accuracy of frequent region clustering against 10 clusterings of binding cavities from individual conformational samples. Phylip used to perform UPGMA clustering on volumetric distances between individual samples. Trees were visualized using Newick Utilities version 1.6 [22].

2.7 Comparing FAVA against statistical models for rigid comparison

We compared FAVA against VASP-S [8], a statistical analysis of protein structures that is trained on original structures without sampled conformations. VASP-S can distinguish variations in original binding cavity shape that are large enough to create different binding preferences from variations that are too small to affect specificity [8]. We hypothesized that the variations we observed between different conformational samples of the same binding cavities were so large that VASP-S would incorrectly classify them as having different binding preferences. These incorrect pre-

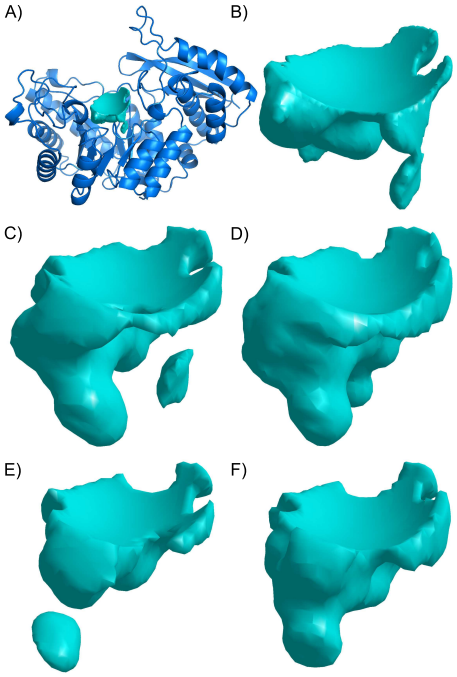

Figure 5: Conformational samples of the ligand binding cavity in yeast enolase (pdb: 1ebh). A) The position of the cavity (teal) within the tertiary structure of enolase (blue cartoon). B) The ligand binding cavity in the original crystal structure. C-F) Binding cavities from other conformational samples of yeast enolase. All panels illustrate the cavity from the same perspective, generated with the global envelope surface, as described earlier.

dictions would demonstrate the importance and utility of a technique like FAVA in enabling accurate comparisons of binding cavities despite the presence of conformational noise.

2.8 Implementation Details

FAVA is a high-level procedure that uses CSG operations from VASP [10]. Computation time was proportional to the volume of the operands: Intersections between amino acids and binding cavities required fractional seconds; operations on molecular surfaces were approximately 10 seconds. Simulations and CSG operations were run on AMD Opteron 6128 processors with 2 gigabytes of memory per core. Figure 5 was generated with custom software and Pymol [12].

3. RESULTS

3.1 Ligand binding cavities vary considerably over time

Figure 5 illustrates changes in the ligand binding cavity of yeast enolase, as sampled from a 100 nanosecond simulation. Sidechain motions, and smaller backbone motions created significant variations that enlarged, shrank, and even separated regions of the cavity. Relative to other proteins in the dataset, the shape of the binding cavity of yeast enolase was not the most variable nor was it the most conserved. Binding cavities in some proteins, such as elastase, varied much more, while others varied less.

In Figure 6, we plot the volumes of binding cavities in

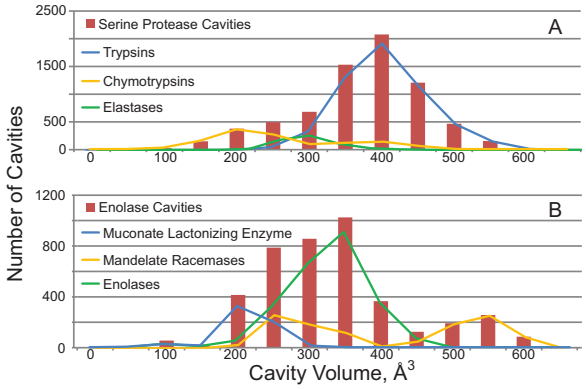

Figure 6: Volume of cavity sizes observed in conformational samples of serine proteases (A) and enolases (B), shown in red bars. Lines plot the quantity and spatial volume of cavities sampled for specific families.

conformational samples of the entire dataset. Among the serine proteases, samples of trypsin cavities ranged from 248 Å^3 to 692 Å^3, chymotrypsin cavities ranged from 276 Å^3 to 568 Å^3, and elastase cavities ranged from 126 Å^3 to 552 Å^3, despite the general principle that chymotrypsin $S1$ cavities are larger to accommodate aromatic sidechains, and elastase cavities are smaller to accommodate amino acids like alanine or valine. Similar variations can be seen amongst the ligand binding cavities of the enolase superfamily. Enolase cavities ranged from 90 Å^3 to 507 Å^3, mandelate racemases ranged from 225 Å^3 to 673 Å^3, and cavities sampled from muconate lactonizing enzyme were between 89 Å^3 and 343 Å^3. This degree of structural variation demonstrates the fundamental difficulty of accurately comparing binding site geometry in the presence of flexibility.

Statistical modeling with a rigid model for classification does not add precision to the structural comparison of flexible binding sites. We used VASP-S [8] to classify all CSG differences between pairs of cavities sampled from the same protein. Over 65 percent of CSG differences were incorrectly classified by VASP-S as being so large as to be consistent with different binding preferences. Based on the frequency of incorrect classifications, a comparison of individual structures has a high probability of being inaccurate.

From these observations, it is clear that the flexibility of serine proteases and enolases creates significant variations between different samples of binding cavities from the same protein. Because of the variability in the data, it is also clear that comparisons of individual conformational samples can point to erroneous similarities and variations that relate to that sample alone and not a larger trend. Thus, a technique like FAVA, which incorporates flexibility from conformational samples into the analysis, is essential for accurate general comparison.

3.2 Evaluating frequent region approximation

FAVA approximates frequent regions using random selections of conformational samples. Actual frequent regions cannot be computed on realistic data because of their combinatorial nature. This situation prevents a direct evaluation of the accuracy of our approximation technique, but it does not prevent us from evaluating the geometric consistency of the approximations generated. Specifically, when

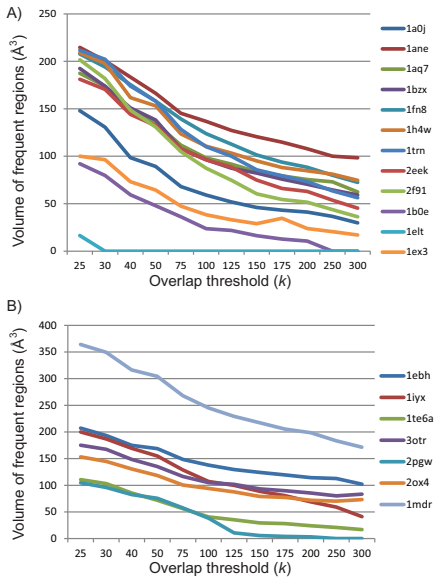

Figure 7: Volumes of frequent regions in serine protease (A) and enolase (B) cavities, computed at varying thresholds.

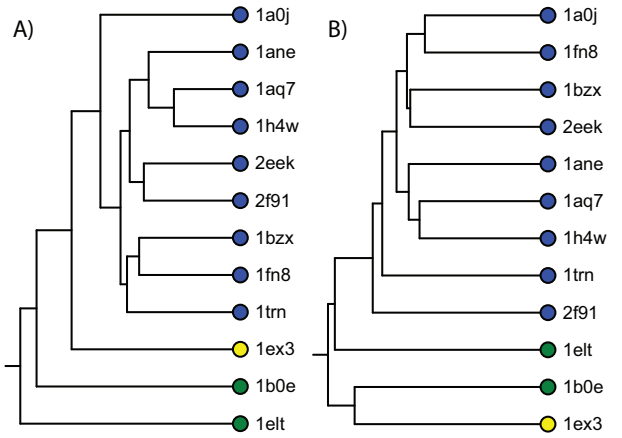

Figure 8: Comparison of clusterings of frequent regions and of individual cavities from serine protease structures. A) Clustering of frequent regions. B) Clustering of cavities from individual conformational samples. In both trees, topology is calculated based on volumetric distance. Coloring, which is independent of clustering topology, indicates the ligand binding preference of the protein.

considering conformational samples from the same protein, frequent regions with higher overlap thresholds must always have equal or smaller volume than frequent regions with lower overlap thresholds. This fact holds logically because regions where k cavities overlap are also, by definition, a region where fewer than k cavities overlap.

We evaluated the degree to which this rule holds for our approximation by computing the volumes of frequent regions at a wide range of overlap thresholds for all proteins in our data set. Figure 7 indicates that volumes of frequent regions are almost monotonically descending as overlap thresholds increase. They also indicate that frequent regions from some proteins remain consistently larger than others, suggesting fewer conformational changes that interfere with the shape of the binding cavity. The only inconsistency appears to be in a small increase in the volume of the approximated frequent regions of crayfish trypsin (pdb: 2f91), at an overlap threshold of 175. It is also notable, though not inconsistent, that volumes of frequent regions from sampled cavities of Atlantic salmon elastase (pdb: 1elt) become zero above overlap thresholds of 25, indicating that conformational flexibility radically alters the shape of that cavity. This effect is shown in detail in Supplementary meterials Section 4. Overall, these observations suggest that FAVA is generating stable, logically consistent approximations of frequent regions.

3.3 Clustering frequent regions

Figure 8a illustrates a UPGMA clustering of serine protease frequent regions based on volumetric distance. Trypsins were correctly clustered away from other serine proteases. Elastases were also separated, but Atlantic salmon elastase was placed as an outlier because it has zero volume. Chymotrypsin was correctly separated from both trypsins and elastases. Figure 8b is an example of a UPGMA clustering generated from randomly selected conformational samples of each protein. We can see that one salmon elastase (pdb: 1elt) is classified as more similar to the trypsins than it is to porcine elastase (pdb: 1b0e), and that 1b0e is more similar to the chymotrypsin than anything else. This kind of miscategorization was typical of other clusteriings of cavities from randomly selected conformational samples.

A UPGMA clustering of frequent regions derived from enolase binding cavities is shown in Figure 9a. Frequent regions from enolase and muconate lactonizing enyzyme were correctly separated, as were frequent regions from mandelate racemase, except that the mandelate racemase from Pseudomonas putida (pdb: 1mdr) was clustered with yeast enolase instead of with mandelate racemase from Zymomonas mobilis (pdb: 2ox4). Clusterings of individual conformational samples of enolase cavities (e.g. Fig 9b) showed similar errors. Overall, UPGMA clustering of frequent regions in the serine proteases and enolases generally reflected differences in specificity and was equal or more accurate than clustering individual conformational samples. This result demonstrates that a flexible representation of binding cavities can diminish classification errors caused by conformational flexibility.

3.4 Influential amino acids

Differences in binding preferences between two proteins can be caused by changes in the backbone and sidechain positions of nearby amino acids. To evaluate how accurately FAVA can detect amino acids that create such changes, we compute the median intersection volume $INT_r(A, B)$, for all residues r in all elastase structures (A), and all non-elastase serine protease cavities (B). For each conformational sample of each elastase residue and each serine protease cavity, we also measured the minimum, 25th percentile, 75th percentile, and maximum volume of intersection.

Most amino acids exhibited zero or very small intersection with any serine protease cavity, including cavities from the same protein, because the amino acid is distant from cavity. Nonetheless, some amino acids do occasionally intersect with binding cavities of the same protein. For example, among amino acids of porcine pancreatic elastase (pdb: 1b0e), the

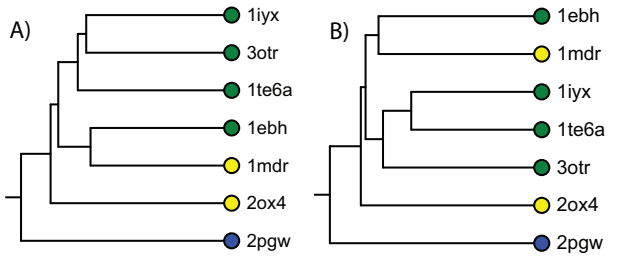

Figure 9: Comparison of clusterings of frequent regions and of individual cavities from enolase structures. A) Clustering of frequent regions. B) Clustering of cavities from individual conformational samples. In both trees, topology is calculated based on volumetric distance. Coloring, which is independent of clustering topology, indicates the ligand binding preference of the protein.

amino acid that most intersects the binding cavity of 1b0e is serine 195, the nucleophilic serine responsible for catalysis in serine proteases [39]. It occupies an median of 5 $Å^3$ inside samples of binding cavities in 1b0e.

When considering intersections between elastase residues and cavities from trypsins, different amino acids exhibited much larger median volumes of intersection. As an example, Figure 10 illustrates the degree of intersection between amino acids of porcine elastase (pdb: 1b0e) and cavities from conformational samples of salmon trypsin (pdb: 1bzx). Samples of valine 216 exhibited a median intersection volume of 45 $Å^3$ with trypsin cavities. Threonine 226 exhibited median intersection volumes of 29 $Å^3$. These predictions correspond to experimental findings: Both V216 and T226 are known to occupy parts of the S1 pocket (inset, Figure 10), shortening it accommodate small hydrophobic residues [43]. We observed similar volumes of intersection between elastase amino acids and other trypsin cavities as well.

Finally, we also measured median intersection volumes between elastase residues and the sampled cavities of bovine chymotrypsinogen (pdb: 1ex3). Again, most amino acids exhibited small or zero median volumes of intersection with cavity samples. Serine 195, valine 216 and threonine 226 exhibited larger median volumes, at 16 $Å^3$, 20 $Å^3$, and 15 $Å^3$, respectively. These results again illustrate that amino acids that alter cavity geometry can be detected despite conformational flexibility in both the amino acids and the cavity.

4. DISCUSSION

We have presented a new volumetric method for the geometric comparison of protein binding cavities. FAVA implements a conformationally general approach to protein structure comparison that permits detailed comparisons of binding cavities despite considerable structural variations. This capability is possible by leveraging a unique representation of molecular flexibility that uses conformational samples to build a detailed picture of how binding cavities frequently vary.

We demonstrated FAVA on applications to the serine protease and enolase superfamilies. Ligand binding cavities in both superfamilies exhibited considerable conformational flexibility. Despite this variability, FAVA was able to classify members of both superfamilies according to known differences in ligand binding preferences. Classifications with

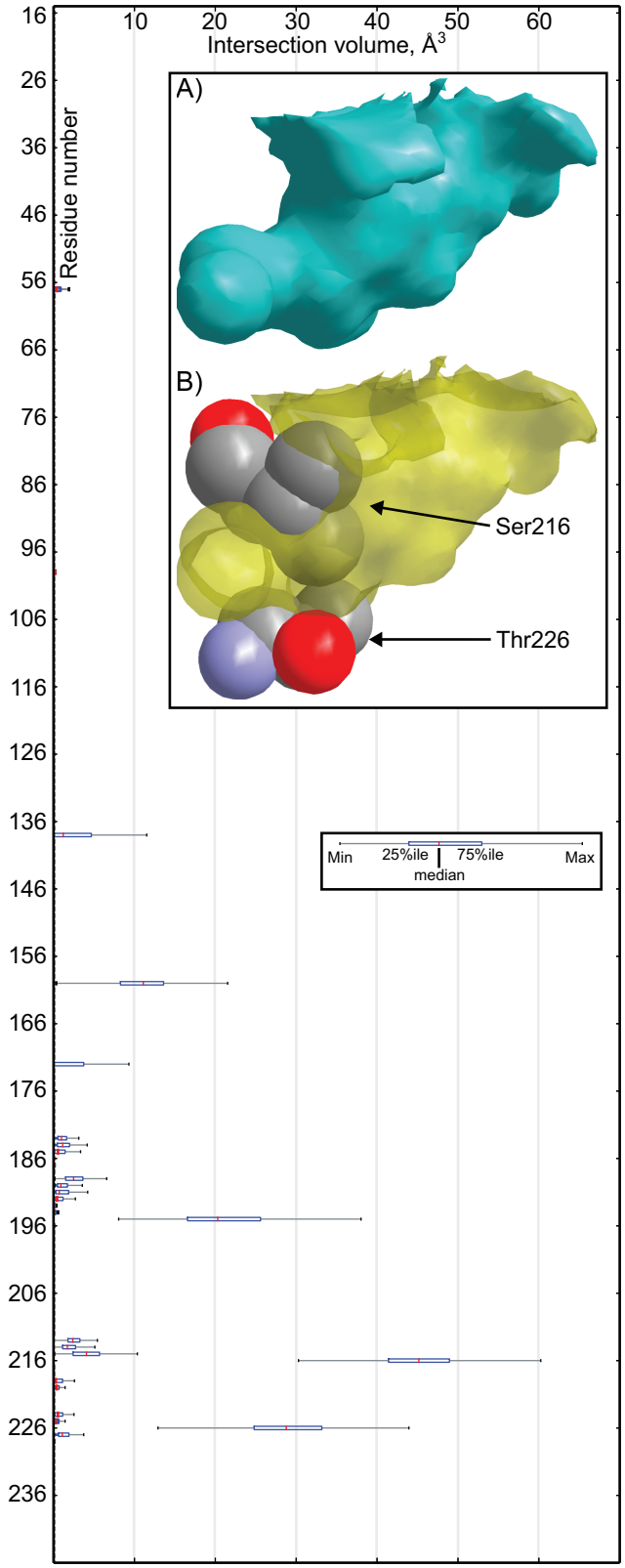

Figure 10: Intersection volume of amino acids from conformational samples of porcine pancreatic elastase (pdb: 1b0e) with cavities from conformational samples of salmon trypsin (pdb: 1bzx). A) The trypsin cavity (teal). B) One snapshot of Val216 and Thr226 from 1b0e, relative to the cavity.

frequent regions were superior to classifications that would be generated if a single conformation had to be selected at random. Measuring the median volume of intersection between sampled amino acids of one protein and the sampled cavities of another, FAVA was also capable of identifying amino acids that have an experimentally established influence on binding specificity, despite the flexibility of their side chains.

As a tool for the flexible volumetric comparison of ligand binding cavities, FAVA has considerable potential for wider applications. First, the precision of FAVA depends on the quality of conformational samples that it are provided. As our capability to simulate molecular conformations expands [41, 2], a larger and more representative range of conformational samples can be provided to FAVA to achieve superior comparison accuracy. Second, in many cases, efforts to create proteins with engineered binding preferences already involve the simulation of protein structures. FAVA introduces an analysis of the resulting simulation data that might yield more detailed comparisons of the binding sites and point to specific amino acids that could be altered for a desired binding preference. Finally, sampled representations of protein structure could also be applied to the detection of remote homologs. By using conformational variations to match distantly related proteins in different conformations, sampled representations might offer an important tools for both function annotation and the analysis of ligand binding specificity.

5. REFERENCES

[1] Babbitt, P. C., Hasson, M. S., Wedekind, J. E., Palmer, D. R., Barrett, W. C., Reed, G. H., Rayment, I., Ringe, D., Kenyon, G. L., and Gerlt, J. A. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. *Biochemistry 35*, 51 (1996), 16489–501.

[2] Beberg, A. L., Ensign, D. L., Jayachandran, G., Khaliq, S., and Pande, V. S. Folding@ home: Lessons from eight years of volunteer distributed computing. In *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on* (2009), IEEE, pp. 1–8.

[3] Berendsen, H., Postma, J., van Gunsteren, W., and Hermans, J. Intermolecular forces. *Pullman, B., Ed.; Reidel Publishing Company: Dordrecht* (1981), 331–342.

[4] Berglund, G., Smalas, A., Outzen, H., and Willassen, N. Purification and characterization of pancreatic elastase from North Atlantic salmon (Salmo salar). *Mol Mar Biol Biotechnol 7*, 2 (1998), 105–14.

[5] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res 28*, 1 (Jan. 2000), 235–42.

[6] Birzele, F., Gewehr, J. E., Csaba, G., and Zimmer, R. VorolignŮfast structural alignment using voronoi contacts. *Bioinformatics 23*, 2 (2007), e205–e211.

[7] Bryant, D. H., Moll, M., Finn, P. W., and Kavraki, L. E. Combinatorial clustering of residue position subsets predicts inhibitor affinity across the human kinome. *PLoS computational biology 9*, 6 (2013), e1003087.

[8] Chen, B., and Bandyopadhyay, S. VASP-S: A Volumetric Analysis and Statistical Model for Predicting Steric Influences on Protein-Ligand Binding Specificity. In *Proceedings of 2011 IEEE International Conference on Bioinformatics and Biomedicine* (2011), pp. 22–9.

[9] Chen, B. Y., Fofanov, V. Y., Bryant, D. H., Dodson, B. D., Kristensen, D. M., Lisewski, A. M., Kimmel, M., Lichtarge, O., and Kavraki, L. E. The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. *J Comp Biol 14*, 6 (2007), 791–816.

[10] Chen, B. Y., and Honig, B. VASP: A Volumetric Analysis of Surface Properties Yields Insights into Protein-Ligand Binding Specificity. *PLoS Comput Biol 6*, 8 (2010), 11.

[11] Connolly, M. Solvent-accessible surfaces of proteins and nucleic acids. *Science 221*, 4612 (Aug. 1983), 709–713.

[12] DeLano, W. L. The PyMOL Molecular Graphics System, 2002.

[13] Dundas, J., Adamian, L., and Liang, J. Structural signatures of enzyme binding pockets from order-independent surface alignment: a study of metalloendopeptidase and nad binding proteins. *Journal of Molecular Biology 406*, 5 (2011), 713–729.

[14] Felsenstein, J. Phylip - phylogeny inference package (version 3.2). 164–166.

[15] Gibrat, J. F., Madej, T., and Bryant, S. H. Surprising similarities in structure comparison. *Curr Opin Struct Biol 6*, 3 (June 1996), 377–85.

[16] Gráf, L., Jancsó, a., Szilágyi, L., Hegyi, G., Pintér, K., Náray-Szabó, G., Hepp, J., Medzihradszky, K., and Rutter, W. J. Electrostatic complementarity within the substrate-binding pocket of trypsin. *Proc Natl Acad Sci U S A 85*, 14 (July 1988), 4961–5.

[17] Gunasekaran, K., and Nussinov, R. How Different are Structurally Flexible and Rigid Binding Sites ? Sequence and Structural Features Discriminating Proteins that Do and Do not Undergo Conformational Change upon Ligand Binding. *J Mol Biol 365* (2007), 257–273.

[18] Hess, B. P-LINCS: a parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation 4*, 1 (Jan. 2008), 116–122.

[19] Hess, B., Kutzner, C., Van Der Spoel, D., and Lindahl, E. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation 4*, 3 (2008), 435–447.

[20] Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation 4*, 3 (Mar. 2008), 435–447.

[21] Holm, L., and Sander, C. Mapping the protein universe. *Science 273*, 5275 (Aug. 1996), 595–603.

[22] Junier, T., and Zdobnov, E. M. The newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics 26*, 13 (2010), 1669–1670.

[23] Konc, J., and Janežič, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics 26*, 9 (2010), 1160–1168.

[24] Kühnel, K., and Luisi, B. F. Crystal structure of the Escherichia coli RNA degradosome component enolase. *J Mol Biol 313*, 3 (Oct. 2001), 583–92.

[25] Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., et al. Clustal w and clustal x version 2.0. *Bioinformatics 23*, 21 (2007), 2947–2948.

[26] Menke, M., Berger, B., and Cowen, L. Matt: local flexibility aids protein multiple structure alignment. *PLoS computational biology 4*, 1 (2008), e10.

[27] Morihara, K., and Tsuzuki, H. Comparison of the specificities of various serine proteinases from microorganisms. *Arch Biochem Biophys 129*, 2 (1969), 620–634.

[28] Mosca, R., and Schneider, T. R. Rapido: a web server for the alignment of protein structures in the presence of conformational changes. *Nucleic acids research 36*, suppl 2 (2008), W42–W46.

[29] Nose, S., and Klein, M. Constant pressure molecular dynamics for molecular systems. *Molecular Physics 50*, 5 (1983), 1055–1076.

[30] Nussinov, R., and Wolfson, H. J. Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc Natl Acad Sci U S A 88*, 23 (Dec. 1991), 10495–9.

[31] Orengo, C. A., and Taylor, W. R. SSAP: Sequential Structure Alignment Program for Protein Structure Comparison. *Method Enzymol 266* (1996), 617–635.

[32] Parrinello, M., and Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics 52* (1981), 7182.

[33] Petrey, D., and Honig, B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Method Enzymol 374*, 1991 (Jan. 2003), 492–509.

[34] Poirrette, a. R., Artymiuk, P. J., Rice, D. W., and Willett, P. Comparison of protein surfaces using a genetic algorithm. *J Comput Aided Mol Des 11*, 6 (Nov. 1997), 557–69.

[35] Rakus, J. F., Fedorov, A. A., Fedorov, E. V., Glasner, M. E., Hubbard, B. K., Delli, J. D., Babbitt, P. C., Almo, S. C., and Gerlt, J. A. Evolution of enzymatic activities in the enolase superfamily: L-rhamnonate dehydratase. *Biochemistry 47*, 38 (Sept. 2008), 9944–54.

[36] Russell, R. B. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol 279*, 5 (June 1998), 1211–27.

[37] Schaer, J., and Stone, M. Face traverses and a volume algorithm for polyhedra. *Lect Notes Comput Sc 555/1991* (1991), 290–297.

[38] Schafer, S. L., Barrett, W. C., Kallarakal, A. T., Mitra, B., Kozarich, J. W., Gerlt, J. A., Clifton, J. G., Petsko, G. A., and Kenyon, G. L. Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the D270N mutant. *Biochemistry 35*, 18 (May 1996), 5662–9.

[39] Schechter, I., and Berger, A. On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications 27*, 2 (1967), 157–162.

[40] Shatsky, M., Nussinov, R., and Wolfson, H. J. FlexProt: alignment of flexible protein structures without a predefinition of hinge regions. *J Comput Biol 11*, 1 (Jan. 2004), 83–106.

[41] Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., et al. Atomic-level characterization of the structural dynamics of proteins. *Science 330*, 6002 (2010), 341–346.

[42] Shindyalov, I. N., and Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng 11*, 9 (Sept. 1998), 739–47.

[43] Shotton, D., and Watson, H. Three-dimensional structure of tosyl-elastase. *Nature 225*, 5235 (1970), 811–816.

[44] Sneath, P. H., and Sokal, R. R. *Numerical taxonomy. The principles and practice of numerical classification.* 1973.

[45] Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence 13*, 4 (1991), 376–380.

[46] Vesterstrøm, J., and Taylor, W. R. Flexible secondary structure based protein structure comparison applied to the detection of circular permutation. *Journal of Computational Biology 13*, 1 (2006), 43–63.

[47] Xie, L., and Bourne, P. E. A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics 8 Suppl 4* (Jan. 2007), S9.

[48] Yang, A.-S., and Honig, B. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J Mol Biol 301*, 3 (Aug. 2000), 665–78.

[49] Ye, Y., and Godzik, A. Multiple flexible structure alignment using partial order graphs. *Bioinformatics 21*, 10 (May 2005), 2362–9.